A Unified hybrid speech-sound generation Framework for Zero-Shot Voice Cloning in Complex Acoustic Scenes

Anonymous ACL submission

Abstract

Synthesizing complex acoustic scenes with zero-shot voice cloning remains a challenge for unified models, primarily due to the control dilemma between representation ambiguity in text-only paradigms and acoustic shortcuts in hybrid conditioning. To address this, we propose ChameleAudio, the first unified framework capable of synchronous speech and sound generation while maintaining high-fidelity zero-shot voice cloning. To resolve the shortcut learning problem, we devise a progressive training strategy. This ordered paradigm prioritizes semantic controllability before refining acoustic details, ensuring the model captures high-level semantic descriptions. Furthermore, to explicitly resolve multi-condition conflicts, we incorporate a Disentangled Flow Matching strategy driven by Independent Condition Masking. By enforcing statistical independence among modalities during training, this mechanism prevents optimization collapse onto the dominant acoustic stream and enables precise multi-directional guidance during inference. Backed by our LLM-driven hybrid data pipeline, ChameleAudio achieves state-of-the-art performance in zero-shot voice cloning within complex acoustic scenes. By effectively balancing speech intelligibility and environmental fidelity, it achieves a WER of 2.65% and an FAD of 5.85. Audio samples are available at https://demoanonymity.github.io/chameleaudio/.

1 Introduction

Real-world auditory scenes are inherently compositional, seamlessly blending specific vocal identities with dynamic environmental acoustics (e.g., a specific person speaking amidst a thunderstorm) (Bregman, 1994). However, despite the rapid progress in generative audio models, current systems remain bifurcated: they either clone voices in isolation (Wang et al., 2023a; Le et al., 2023; Tan et al., 2024) or generate generic sounds (Liu et al., 2023; Kreuk et al., 2023; Ghosal et al., 2023), failing to synthesize high-fidelity, synchronous audio scenes where customized speech and complex environments coexist. The inability of existing methods to achieve such unified generation stems from three deep-seated limitations involving representation, optimization, and data.

Despite the evolution of diverse text conditioning strategies, ranging from semantic captions to complex instructions, the prevailing 'text-only' paradigm is still fundamentally bottlenecked by representation ambiguity. While instruction-driven unified frameworks (Qiang et al., 2025; Vyas et al., 2023; Yang et al., 2023) and diffusion-based TTA models (Liu et al., 2023, 2024; Ghosal et al., 2023; Kreuk et al., 2023) successfully utilize natural language for generic generation, they face an inherent "one-to-many" mapping problem in speech synthesis. Abstract textual descriptions remain acoustically underspecified compared to acoustic prompt-based systems (Le et al., 2023; Tan et al., 2024; Wang et al., 2023a). Consequently, these text-only unified frameworks are fundamentally incapable of performing high-fidelity zero-shot voice cloning—the task of replicating a unique vocal identity from a brief audio prompt. This creates a critical capability gap, leaving them far behind specialized zero-shot TTS models (Du et al., 2024; Anastassiou et al., 2024) in personalization and speaker identity preservation.

To overcome representation ambiguity and achieve precise multi-condition control, hybrid conditioning integrates acoustic prompts alongside textual instructions. However, this paradigm faces a critical bottleneck: *acoustic shortcut learning*, stemming from the inherent information asymmetry between modalities. While this integration aims for unified control, it frequently precipitates *modality collapse* (Daunhawer et al., 2021). Specifically, since acoustic references possess significantly higher information density than sparse

textual descriptions, the model tends to prioritize the dominant acoustic stream during optimization. This tendency leads the model to exploit the reference audio as a "shortcut" to minimize the generation objective (Geirhos et al., 2020), thereby bypassing the intended semantic guidance. Consequently, the effectiveness of descriptive text controls is compromised, limiting the model's capability to flexibly adhere to semantic instructions beyond the acoustic reference.

Moreover, the absence of high-quality synchronous data significantly impedes the advancement of synchronous speech and sound generation. Current audio datasets remain isolated: speech synthesis relies on clean corpora like LibriTTS (Zen et al., 2019), while audio generation uses event-specific datasets like AudioSet (Gemmeke et al., 2017). The absence of semantically aligned mixtures limits models' ability to learn the joint distribution and physical interactions between vocals and dynamic environments.

To tackle the challenge of zero-shot voice cloning within complex acoustic scenes, we present ChameleAudio, a unified multi-modal diffusion transformer framework designed for high-fidelity, synchronous speech and sound synthesis. To mitigate multi-modal interference and acoustic shortcuts, we introduce: (1) PL: a Progressive Learning strategy that orders the training into a logical sequence, prioritizing standalone semantic generation before incrementally learning timbre cloning; and (2) DFM: a Disentangled Flow Matching strategy driven by Independent Condition Masking (ICM). Unlike rigid geometric projections, this approach enforces statistical independence between modalities during training, compelling the model to learn robust marginal distributions and preventing optimization collapse onto the dominant acoustic stream. To support this framework, we constructed a novel Multi-Task Compositional Audio Dataset designed for multi-condition supervision, facilitating robust joint modeling.

Our key contributions are as follows:

- We propose ChameleAudio, the first unified framework for zero-shot voice cloning within complex acoustic scenes, seamlessly integrating speech and sound synthesis to achieve unified generation.
- We introduce a Disentangled Flow Matching strategy via Independent Condition Masking. This mechanism resolves the acoustic shortcut problem by enforcing statistical independence among modalities, enabling flexible multi-directional guidance for precise semantic control during inference.
- We develop a scalable Automated Data Pipeline to overcome the lack of semantically coherent hybrid data, enabling the model to learn the joint distribution of speech and sound effects.
- Experiments show that ChameleAudio achieves state-of-the-art performance in zero-shot voice cloning within complex acoustic scenes.

2 ChameleAudio

2.1 Architecture

ChameleAudio employs a Multi-Modal Diffusion Transformer (MM-DiT) backbone operating on a unified latent sequence. We formalize the processing of heterogeneous inputs into a joint high-dimensional space ($D = 1024$).

Input Representation and Unification. To effectively capture semantic, linguistic, and acoustic features, we formulate the model conditioning into four distinct streams before unification. First, the *Instruction Stream* ($\mathbf{H}_{inst}$) is encoded by Qwen and projected to $\mathbb{R}^{B\times L_I\times D}$, providing global semantic guidance on environmental atmosphere and speaker attributes. Second, the *Text Stream* ($\mathbf{H}_{txt}$) processes linguistic content via a Zipformer (Zhu et al., 2025) encoder, yielding phonetic embeddings for speech data or special effect tokens for sound data. Third, to inject vocal identity, we extract a global Speaker Embedding $\mathbf{e}_{spk} \in \mathbb{R}^{B\times D_{spk}}$ from the *Reference Audio Stream* using the pre-trained CAM++ (Wang et al., 2023b) encoder. Finally, the *Target Audio Stream* ($\mathbf{x}_t$) is compressed by Mel-VAE (Wang et al., 2025a) into continuous latents to serve as the diffusion input.

Stream Fusion and Sequence Modeling. To synthesize audio that respects both linguistic content and speaker identity, we first integrate the text and speaker streams into a fused representation. Let $\mathbf{e}_{spk} \in \mathbb{R}^{B\times D}$ be the global speaker embedding. For non-speech samples (Type B), $\mathbf{e}_{spk}$ is set to a **fixed zero vector** $\mathbf{0} \in \mathbb{R}^D$ to deactivate identity-specific pathways, ensuring the model relies solely

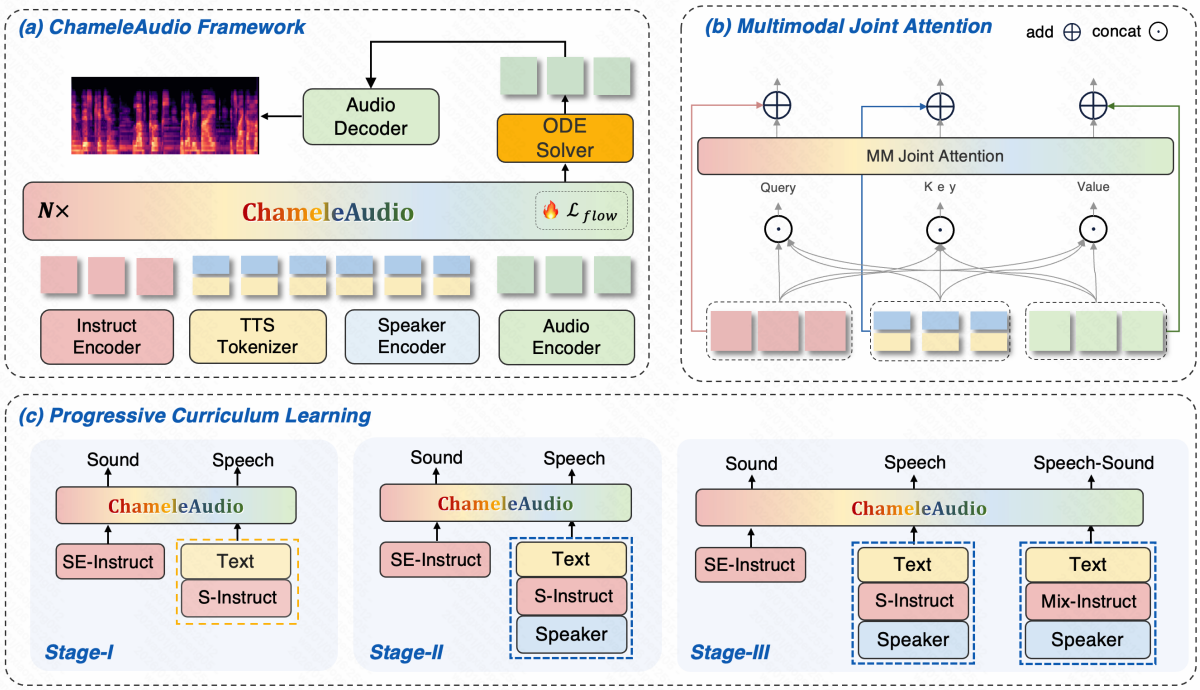

Figure 1: The ChameleAudio framework, illustrated in (a), integrates multi-modal encoders with a Flow Matching-based generator to process Instruction, Text, Speaker, and Audio streams; (b) details the Multimodal Joint Attention mechanism designed for unified sequence interaction; finally, (c) depicts the Progressive Curriculum Learning strategy, advancing from basic generation via SE-Instruct and S-Instruct in Stage-I, to zero-shot cloning in Stage-II, and culminating in compositional mixed audio generation using Mix-Instruct in Stage-III.

on environmental descriptions. The fused representation $\mathbf{H}_{fused} \in \mathbb{R}^{B\times L_T\times D}$ is formulated as:

$$\mathbf{H}_{fused} = \mathcal{F}_{fuse}\left([\mathbf{H}_{txt} \parallel \text{Tile}(\mathbf{e}_{spk}, L_T)]\right) \quad (1)$$

where $\parallel$ denotes channel-wise concatenation, and $\text{Tile}(\cdot, L_T)$ replicates the global vector to match the phonetic sequence length.

To facilitate cross-modal interaction without hard-coded alignment priors, we unify the heterogeneous streams into a single high-dimensional sequence $\mathbf{Z}$. By temporally concatenating the three streams, we define the input to the transformer backbone as:

$$\mathbf{Z}_0 = [\mathbf{H}_{inst}; \mathbf{H}_{fused}; \mathbf{x}_t] \in \mathbb{R}^{B\times L\times D} \quad (2)$$

where L is the aggregate sequence length. This sequence is processed by N stacked MM-DiT blocks. Within each block, an MM Joint Attention mechanism computes global dependencies across all modality tokens simultaneously:

$$\mathbf{Z}_{l+1} = \mathbf{Z}_l + \text{MHA}(\text{LN}(\mathbf{Z}_l)) \quad (3)$$

As illustrated in Figure 1(b), this formulation allows the audio latents to dynamically query semantic cues in $\mathbf{H}_{inst}$ and phonetic structures in $\mathbf{H}_{fused}$ without explicit modal boundaries, effectively learning complex alignment and generation rules in a purely data-driven manner.

2.2 Mask-Driven Disentanglement

To address the acoustic shortcut problem, we propose an Independent Condition Masking strategy. Unlike standard joint optimization, this approach enforces statistical independence among modalities, enabling disentangled gradient control during inference.

Training: Independent Dropout. Let $\mathcal{C} = \{\mathbf{c}_{spk}, \mathbf{c}_{txt}, \mathbf{c}_{inst}\}$ denote the full condition set. During training, we apply a stochastic masking operation $\mathcal{M}$ to ensure the model learns the marginal distributions of each modality. We sample a binary mask vector $\mathbf{m} \in \{0,1\}^3$ from a Bernoulli distribution $\mathcal{B}(p_{drop})$ independently for each stream. The masked condition set $\tilde{\mathcal{C}}$ is defined as:

$$\tilde{\mathbf{c}}_k = \begin{cases} \mathbf{c}_k & \text{if } m_k = 1 \\ \varnothing & \text{if } m_k = 0 \end{cases}, \quad \forall k \in \{spk, txt, inst\} \quad (4)$$

where $\varnothing$ represents a learnable null embedding. The flow matching objective is then optimized conditioned on $\tilde{\mathcal{C}}$. This independence assumption pre-

vents the optimization trajectory from collapsing onto the dominant acoustic manifold, ensuring robust generation capabilities even under partial conditioning.

Inference: Multi-Directional Guidance. Leveraging the independently trained conditions, we employ *Multi-Condition Classifier-Free Guidance (CFG)* to explicitly isolate semantic and acoustic gradients. The predicted velocity field v_{pred} is formulated as a linear combination of the unconditional estimate and disentangled conditional deviations:

$$v_{pred}(\mathbf{z}_t) = v_\theta(\mathbf{z}_t, \varnothing) + \sum_{k \in \mathcal{C}} s_k \cdot \Delta v_k(\mathbf{z}_t)$$
$$\text{where} \quad \Delta v_k(\mathbf{z}_t) = v_\theta(\mathbf{z}_t, \mathbf{c}_k) - v_\theta(\mathbf{z}_t, \varnothing) \tag{5}$$

Here, s_k represents the guidance scale for condition k. This formulation allows for precise rebalancing of modality influence; specifically, increasing s_{inst} relative to s_{spk} amplifies semantic adherence while suppressing acoustic shortcuts, effectively resolving the interference between entangled representations.

2.3 Progressive Training Strategy

Training a unified model on heterogeneous modalities poses gradient conflict challenges. We propose a three-stage progressive training, utilizing Conditional Flow Matching (CFM) as the objective:

$$\mathcal{L}_{CFM}(\theta) = \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_1}\left[||v_\theta(t, \mathbf{Z}) - (\mathbf{x}_1 - \mathbf{x}_0)||^2\right] \tag{6}$$

Stage-I: Semantic Structure Alignment. We mask the Speaker Embedding ($\mathbf{e}_{spk} = \mathbf{0}$). Conditioned solely on Qwen semantic features and phonemes, the model operates on pure speech and sound data via optimal transport paths. The objective is to establish a mapping from semantic descriptions to generalized audio textures, enabling the model to understand the fundamental correspondence between environments and speech attributes (e.g., emotion, gender) without identity constraints.

Stage-II: Acoustic Identity Injection. We unmask $\mathbf{e}_{spk}$ and fine-tune on the pure speech subset. The model learns to utilize CAM++ features to anchor the generated waveforms to specific speaker timbres. This stage bridges the gap between generic text-to-speech and personalized voice cloning, effectively resolving the representation ambiguity inherent in textual descriptions.

Stage-III: Compositional Universal Fusion. The model undergoes final fine-tuning on the full Multi-Task Compositional Audio Dataset. By exposing the model to superimposed samples, it learns the physical interaction rules between speech and background sounds. The CFM objective function ensures the stability of generation trajectories within complex distributions, preserving background atmosphere without compromising speech intelligibility.

Category	Duration (h)	Primary Focus
Type A	80,000	Speaker Identity & Prosody
Type B	10,000	Environmental Textures
Type C	8,000	Joint Distribution & SNR Robustness
Total	98,000	Bilingual (Ch & En)

Table 1: Statistics of the Multi-Task Compositional Audio Dataset.

3 Data Construction

To bridge the data gap in unified audio modeling, we establish a scalable automated pipeline to construct the Multi-Task Compositional Audio Dataset, a high-fidelity, semantic-aligned corpus totaling approximately 98,000 hours. As illustrated in Figure 2, we employ a Structured Natural Language Prompting strategy to organize samples based on explicit semantic indicators within the instructions.

We utilize a proprietary audio-caption model to generate dual-stream annotations : *Text Data* (phonetic transcripts) and *Instruction Data* (structured descriptions). Crucially, the Instruction Data embeds Semantic Content Indicators to define the modality, followed by attribute-specific metadata. Based on these indicators, the dataset is categorized into three distinct types:

Type A: Pure Speech. This subset focuses on clean vocal generation. Instructions explicitly affirm the presence of speech while negating sound effects (e.g., "This audio contains speech. This audio does not contain sound effect"). It provides diverse supervision across fine-grained attributes including gender, age, emotion, and style, paired with accurate phonetic transcripts.

Type B: Pure Sound. Designed for environmental audio tasks, this category emphasizes acoustic scene and event textures. To inhibit phoneme processing, instructions explicitly negate speech (e.g., "This audio does not contain speech."), while the

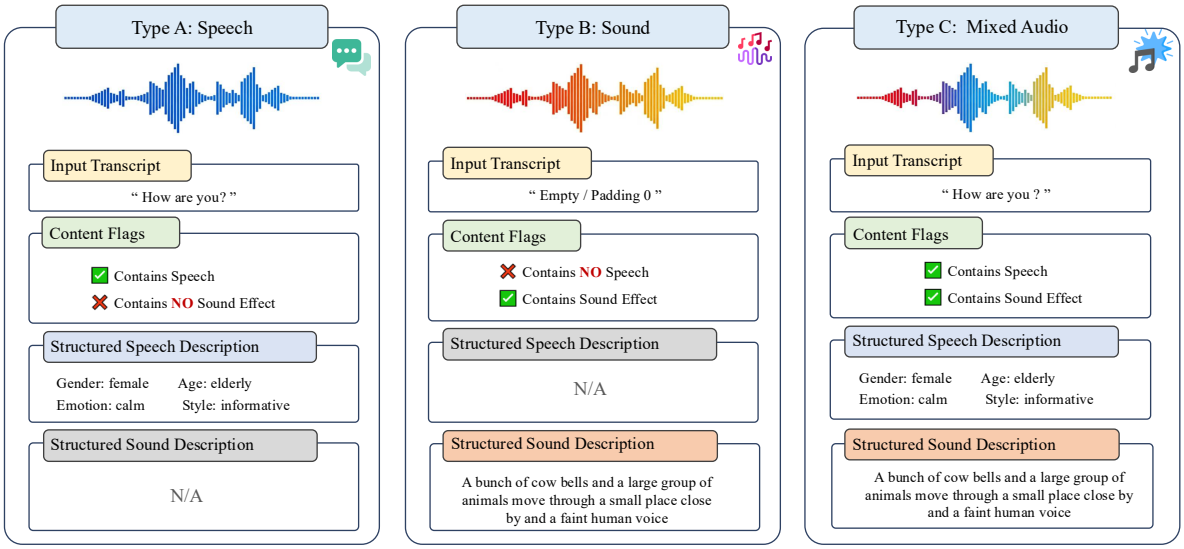

Figure 2: (a) Type A: Speech: Instructions specify speech-only generation, with detailed attributes and text transcripts.(b) Type B: Sound effects: Instructions focus on speech-free content, using effect tokens and describing acoustic scenes or events. (c) Type C: Mixed audio: Instructions activate both speech and sound effects.

text field is occupied by special effect tokens (e.g., "[sound effect]").

Type C: Mixed Audio. This category captures the joint distribution of vocals and environments by activating both modalities (e.g., "This audio contains speech. This audio contains sound effect"). It integrates real-world recordings with synthetic mixtures generated via waveform superposition: $y_{mix} = y_s + \lambda \cdot y_e$. Here, the speech signal y_s serves as the anchor, with the scaling factor λ adaptively adjusted to reach target Signal-to-Noise Ratios (SNR) of 5, 10, and 15 dB. This multi-level SNR strategy enables the model to generalize across varying acoustic complexities.

4 Experiments

4.1 Implementation Details

Model Configuration. ChameleAudio is instantiated with approximately 1.5 billion parameters, featuring a flow matching backbone with a hidden dimension of 1024 and 16 Joint Diffusion Transformer layers equipped with Rotary Positional Embeddings (RoPE) (Su et al., 2024). To facilitate the Independent Condition Masking strategy, we employ a condition dropout ratio of 0.2 during training. For inference, we adopt the Euler ODE solver with 32 Number of Function Evaluations (NFE). For conditioning, we utilize the pre-trained Qwen2.5-Omni (Team et al., 2024) to extract high-level semantic representations and a Zipformer-based (Zhu et al., 2025) phoneme encoder with a feedforward dimension of 512 for linguistic features. The audio processing pipeline employs a Mel-VAE encoder consistent with the Kling–Foley (Wang et al., 2025a) architecture, which compresses 44.1 kHz waveforms into 40-dimensional latent embeddings at a frame rate of 43 Hz, achieving an effective 1024× temporal downsampling. Finally, we employ BigVGAN (Lee et al., 2022) as the vocoder to reconstruct high-fidelity audio waveforms.

4.2 Evaluation Benchmarks and Metrics

To comprehensively assess ChameleAudio, we conduct a two-tiered evaluation strategy covering both fundamental single-modal generation and complex compositional scene synthesis.

Single-Modal Benchmarks. We first benchmark the model on isolated tasks to verify that unified modeling maintains high performance on specific modalities. (1) *Zero-Shot TTS:* We evaluate voice cloning capabilities using the Seed-TTS test set (Anastassiou et al., 2024), which includes 1,088 samples from Common Voice (English) (Ardila et al., 2020) and 2,020 samples from DiDiSpeech (Chinese) (Guo et al., 2021). We measure intelligibility via Word Error Rate (WER) using FunASR (Gao et al., 2023) and quantify speaker identity preservation using Speaker Similarity (SIM), calculated as cosine similarity between WavLM-large-based (Chen et al., 2022) embeddings of the generated speech and the reference prompt. (2)

Text-to-Audio (TTA): We perform evaluation on the AudioCaps test set (Kim et al., 2019). We report standard objective metrics including Fréchet Audio Distance (FAD) (Kilgour et al., 2018) for distribution fidelity, Kullback–Leibler Divergence (KL) for instance quality, and CLAP Score (Elizalde et al., 2023) for text-audio alignment. (3) *Instruction Controllability:* We curated a specialized Instruction Control Test Set comprising 500 samples, each annotated with explicit semantic constraints. We assess the model's adherence to these instructions by calculating the classification accuracy for Gender, Age, Emotion, and Style using pre-trained audio attribute classifiers.

Compositional Generation Benchmarks. To assess the model's capability in synthesizing complex scenes, we curated an Explicit Compositional Test Set comprising 1,000 paired samples, formed by combining speech samples from the Seed-TTS test set (500 English and 500 Chinese) with environmental sound samples from the AudioCaps test set. For evaluation, we directly apply standard single-modal metrics to the generated composite audio to measure the quality of both speech and environmental components within the mixed scene.

Subjective Evaluation. To validate real-world perceptual quality beyond objective metrics, we recruited 15 professional evaluators to perform blind testing on randomly sampled outputs. All ratings follow a standard 5-point Likert scale. (1) For pure sound generation (TTA), evaluators rate the Overall Quality (OVL), assessing audio fidelity and clarity. (2) For speech synthesis (TTS), we report the Mean Opinion Score (MOS) focused on speech naturalness and prosody. (3) For compositional generation, we introduce a Scene-Speech Consistency (SSC) metric, where evaluators rate the degree of semantic matching between the vocal style and the environmental atmosphere.

4.3 Comparison with Existing Method

Performance on Single-Modal Generation. Table 2 presents a unified comparison of control capabilities and quantitative metrics. ChameleAudio distinguishes itself by supporting a comprehensive set of capabilities—ranging from gender and emotion to zero-shot timbre cloning—within a single framework.

In the TTS domain, our model demonstrates exceptional zero-shot voice cloning performance. It achieves WER scores of 1.48% (EN) and 1.29% (ZH), surpassing specialized NAR baselines such as F5-TTS (1.89%/1.53%) (Chen et al., 2025) and ZipVoice (1.64%/1.70%) (Zhu et al., 2025), while remaining competitive with larger models like MaskGCT (Wang et al., 2024). A critical advantage lies in identity control: while unified baselines like InstructAudio (Qiang et al., 2025) are limited to text-based attributes, ChameleAudio leverages acoustic references for precise timbre cloning, filling a crucial gap in unified modeling. Subjective evaluations further confirm this, where our model achieves a quality score of 4.03, demonstrating high perceptual naturalness comparable to specialized baselines like CosyVoice2 (4.02).

Regarding TTA performance, ChameleAudio yields an FAD of 4.47 and a CLAP score of 0.22, delivering generation quality comparable to foundational models like AudioLDM-L (FAD 4.32) (Liu et al., 2023). While there remains a gap compared to specialized TTA SOTA models (Hung et al., 2024)), this reflects a deliberate design priority: our framework is optimized to maintain high-fidelity vocal identity and enable synchronous speech-sound generation, rather than solely maximizing environmental texture metrics. We argue that this trade-off is acceptable, as the current audio quality is sufficient to serve as background atmosphere for complex auditory scenes, a capability that specialized TTA models lack entirely. This is supported by a subjective score of 3.76, indicating that the generated environmental textures remain perceptually satisfactory for background atmosphere.

Performance on Compositional Speech-Sound Generation. The core advantage of ChameleAudio lies in its ability to synthesize spectrally and semantically compatible speech-sound mixtures. To rigorously evaluate this, we compare our unified approach against Cascade Baselines using varying mixing strategies. These range from Direct Superposition (linear addition without adjustment) to SNR-controlled Mixing at 5, 10, and 15 dB, where the speech signal serves as the fixed anchor and the environmental sound is dynamically scaled to match the target Signal-to-Noise Ratio.

The results in Table 2 reveal an inherent trade-off in cascade systems between speech intelligibility and environmental presence. At low SNR levels (5dB) or Direct Mix, cascade models suffer from catastrophic degradation in intelligibility due to auditory masking, with WERs spiking to 8.50% and

Type	Model	Param	Control Capabilities				TTS Metrics				TTA Metrics					Subj.
			G&A	E&S	Spk	TTA	WER(%)↓ EN	WER(%)↓ ZH	SIM↑ EN	SIM↑ ZH	FAD↓	FD↓	KL↓	IS↑	CLAP↑	
TTA	AudioLDM-L (Liu et al., 2023)	739M	✗	✗	✗	✓	-	-	-	-	4.32	29.50	1.68	8.17	0.21	3.45±0.12
TTA	Tango-FT (Ghosal et al., 2023)	866M	✗	✗	✗	✓	-	-	-	-	2.68	15.64	**1.24**	8.78	0.29	-
TTA	EzAudio-XL (Hai et al., 2024)	875M	✗	✗	✗	✓	-	-	-	-	3.64	14.98	1.29	11.38	0.31	-
TTA	Stable Audio (Evans et al., 2025)	1.0B	✗	✗	✗	✓	-	-	-	-	4.19	39.14	2.36	10.07	0.21	3.55±0.15
TTA	TangoFlux (Hung et al., 2024)	516M	✗	✗	✗	✓	-	-	-	-	2.41	20.65	1.27	**12.81**	**0.32**	4.08±0.08
TTA	GenAU-L (Haji-Ali et al., 2024)	1.2B	✗	✗	✗	✓	-	-	-	-	**2.07**	**14.58**	1.36	10.43	0.30	**4.12**±0.07
TTA	**ChameleAudio (TTA)**	1.5B	✓	✓	✓	✓	-	-	-	-	4.47	38.01	2.24	6.71	0.22	3.76±0.10
TTS	MaskGCT (Wang et al., 2024)	1.0B	✗	✗	✓	✗	2.26	2.40	**0.71**	**0.77**	-	-	-	-	-	-
TTS	F5-TTS (Chen et al., 2025)	0.3B	✗	✗	✓	✗	1.89	1.53	0.66	0.75	-	-	-	-	-	4.05±0.10
TTS	Zipvoice(Zhu et al., 2025)	100M	✗	✓	✓	✗	1.64	1.70	0.70	0.75	-	-	-	-	-	**4.10**±0.09
TTS	CosyVoice2 (Du et al., 2024)	0.3B	✗	✓	✓	✗	2.57	1.45	0.65	0.75	-	-	-	-	-	4.02±0.08
TTS	M3-TTS (Wang et al., 2025b)	0.3B	✗	✓	✓	✗	**1.36**	1.31	0.60	0.62	-	-	-	-	-	-
TTS	InstructAudio (Qiang et al., 2025)	1.3B	✓	✓	✗	✓	1.52	1.35	-	-	-	-	-	-	-	-
TTS	**ChameleAudio (TTS)**	1.5B	✓	✓	✓	✓	1.48	**1.29**	0.62	0.69	-	-	-	-	-	4.03±0.12
Joint	Cascade (Direct Mix)	-	-	-	-	-	43.09	40.59	0.44	0.48	10.76	62.59	**1.25**	3.23	**0.26**	1.95±0.25
Joint	Cascade (SNR=5dB)	-	-	-	-	-	8.50	7.80	0.49	0.51	**4.20**	**35.10**	2.10	**6.80**	0.21	3.10±0.20
Joint	Cascade (SNR=10dB)	-	-	-	-	-	4.40	4.21	0.56	0.57	9.48	73.64	1.67	2.45	0.19	3.45±0.18
Joint	Cascade (SNR=15dB)	-	-	-	-	-	**2.10**	**1.90**	**0.60**	0.62	19.38	78.56	1.92	2.23	0.17	3.65±0.16
Joint	**ChameleAudio (Joint)**	1.5B	✓	✓	✓	✓	2.65	2.45	**0.60**	**0.64**	5.85	42.50	2.42	4.45	0.21	**4.15**±0.09

Note: **G&A** = Gender & Age, **E&S** = Emotion & Style, **Spk** = Zero-shot Voice Cloning, **TTA** = Text-to-Audio. **Subj.**: Subjective Quality Score (Estimated based on objective metrics correlation). Models with "-" in Subj. are closed-source or unavailable for subjective testing. Best results are bolded.

Table 2: Unified comparison of Single-Modal and Compositional Generation performance. **Joint** refers to the compositional generation task. Best results are bolded.

43.09%, respectively. Conversely, while increasing the SNR to 15dB recovers speech clarity (WER 2.10%), it severely compromises the richness of the environmental texture, as evidenced by the sharp increase in FAD to 19.38.

In contrast, ChameleAudio effectively breaks this zero-sum trade-off. It achieves robust speech intelligibility (WER 2.65% EN / 2.45% ZH) comparable to the clean 15dB cascade baseline, while simultaneously delivering superior environmental fidelity. Notably, our model achieves an FAD of 5.85, significantly outperforming even the loudest Direct Mix baseline (FAD 10.76). This indicates that instead of simple volume adjustment, ChameleAudio learns the joint distribution of vocals and environment, automatically performing spectral ducking to maintain speech clarity without sacrificing the richness of the acoustic scene. Furthermore, the high Speaker Similarity scores (0.60/0.64) confirm that our model preserves vocal identity integrity even within complex, high-fidelity auditory environments. In terms of subjective quality, evaluators rated ChameleAudio significantly higher (4.15) than the best cascade baseline (3.65), favoring its seamless spectral integration.

4.4 Effect of Progressive Training Strategy

Table 3 presents the ablation analysis of the training strategy, where the One-Stage baseline incorporates the same ACD mechanism to isolate the impact of the curriculum schedule. The results indicate that simultaneous optimization of heterogeneous modalities in the One-Stage setting leads to sub-optimal convergence, evidenced by inferior speech intelligibility and environmental fidelity compared to the final model. Tracking the performance trajectory across stages reveals the necessity of the proposed ordered paradigm. In Stage 1, the model focuses exclusively on semantic-to-acoustic mapping, achieving the highest environmental fidelity with an FAD of 4.42. Upon introducing speaker embeddings in Stage 2, we observe a peak in identity preservation with a SIM-ZH of 0.71; however, this comes at the cost of catastrophic forgetting, where environmental texture quality degrades significantly to an FAD of 5.80. The final Stage 3 effectively resolves this conflict by reintroducing mixed data, recovering environmental generation capabilities to an FAD of 4.47 while yielding the best speech intelligibility. Although the SIM score experiences a minor reduction compared to Stage 2, this is partially attributable to the robustness limitations of the pre-trained speaker encoder, which inadvertently captures environmental noise

Part I: Speech Synthesis Performance				
Setting	**WER↓**		**SIM↑**	
	EN	ZH	EN	ZH
One-Stage	1.59	1.45	0.57	0.63
Stage-I	1.80	1.65	-	-
Stage-II	1.53	1.35	**0.63**	**0.71**
Stage-III	**1.48**	**1.29**	0.62	0.69
Part II: Sound Generation Quality					
Setting	**FAD↓**	**FD↓**	**KL↓**	**IS↑**	**CLAP↑**
One-Stage	5.30	43.50	2.58	6.12	0.18
Stage-I	**4.42**	**37.90**	**2.22**	**6.75**	0.21
Stage-II	5.80	49.20	2.95	5.50	0.15
Stage-III	4.47	38.01	2.24	6.71	**0.22**

Table 3: Quantitative ablation of the training stages. Part I evaluates speech performance (WER, SIM), noting that SIM is N/A for Stage-I due to masked identity.

Part I: Speech Control Accuracy					
Stage	**Setting**	**Gender↑** (%)	**Age↑** (%)	**Emo↑** (%)	**Style↑** (%)
Stage-II	w/o ICM	100.0	84.5	79.2	60.5
	w/ ICM	**100.0**	**86.7**	**81.2**	**62.2**
Stage-III	w/o ICM	100.0	83.8	80.4	68.8
	w/ ICM	**100.0**	**85.2**	**82.5**	**71.5**
Part II: Sound Generation Quality						
Stage	**Setting**	**FAD↓**	**FD↓**	**KL↓**	**IS↑**	**CLAP↑**
Stage-II	w/o ICM	-	-	-	-	-
	w/ ICM	5.80	49.2	2.95	5.50	0.15
Stage-III	w/o ICM	4.89	41.5	2.45	6.42	0.19
	w/ ICM	**4.47**	**38.0**	**2.24**	**6.71**	**0.22**

Table 4: Ablation analysis of the ICM. Part I reports classification accuracies for speech attributes. Part II reports acoustic metrics for sound generation quality.

alongside vocal identity when processing mixed signals, resulting in slightly entangled embeddings. Nevertheless, the overall performance confirms that the progressive strategy is essential for balancing high-frequency speech harmonics and stochastic environmental textures within a unified framework.

4.5 Effect of Independent Condition Masking

Table 4 presents the ablation study on the ICM strategy, offering empirical insights into the acoustic shortcut phenomenon. The results in Part I reveal a distinct behavioral pattern: while Gender accuracies remain at a perfect 100.0% across all settings, indicating that low-level biological cues are robustly captured from the acoustic reference, the control over fine-grained semantic attributes significantly deteriorates without ICM. Specifically, in Stage-III, removing the mechanism causes Age, Emotion, and Style accuracies to drop to 83.8%, 80.4%, and 68.8% respectively. This suggests that without the statistical independence enforced by random masking, the model optimizes for the joint distribution dominated by the information-dense acoustic stream, thereby neglecting sparse instructions regarding style and emotion. This dependency also negatively impacts environmental generation, as shown in Part II. In the absence of ICM, Stage-III exhibits degraded environmental fidelity (FAD 4.89) and reduced text-audio alignment (CLAP 0.19). By re-introducing ICM, the model learns to respect the marginal distributions of each modality. This enables precise multi-directional guidance during inference, effectively amplifying semantic control to recover performance in both fine-grained speech attributes and high-fidelity sound generation.

5 Conclusion

We introduced ChameleAudio, a unified diffusion transformer for high-fidelity compositional audio synthesis. By synergizing a Disentangled Flow Matching strategy driven by ICM with a tailored Progressive Training curriculum, our approach effectively overcomes modality collapse and acoustic shortcuts. Supported by our novel Multi-Task Compositional Audio Dataset, the model achieves state-of-the-art performance in generating complex scenes where customized speech and dynamic environments coexist. Future directions include optimizing inference latency and exploring long-form narrative generation.

6 Limitations

Despite promising results, our work faces three main limitations. First, as a diffusion-based model with 1.5B parameters, inference latency remains a bottleneck for real-time deployment. Second, reliance on automated labeling introduces potential semantic noise, which may propagate hallucinations to the generation process. Third, the model is currently optimized for short segments; maintaining long-form consistency over extended narratives requires further architectural exploration.

7 Ethics Statement

To mitigate risks associated with zero-shot voice cloning (e.g., deepfakes), we enforce strict safety

measures. We plan to incorporate imperceptible watermarking for provenance tracking and adopt a restricted release policy, granting model access only to verified researchers. Furthermore, all training data has undergone rigorous anonymization to protect privacy. We strictly condemn malicious misuse and advocate for responsible AI development.

References

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, and 1 others. 2024. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference*, pages 4218–4222.

Albert S Bregman. 1994. *Auditory scene analysis: The perceptual organization of sound*. MIT press.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. 2025. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6255–6271.

Imant Daunhawer, Thomas M Sutter, Kieran Chin-Cheong, Emanuele Palumbo, and Julia E Vogt. 2021. On the limitations of multimodal vaes. *arXiv preprint arXiv:2110.04121*.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, and 1 others. 2024. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. 2025. Stable audio open. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, Zhangyu Xiao, and 1 others. 2023. Funasr: A fundamental end-to-end speech recognition toolkit. *arXiv preprint arXiv:2305.11013*.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE.

Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. 2023. Text-to-audio generation using instruction guided latent diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3590–3598.

Tingwei Guo, Cheng Wen, Dongwei Jiang, Ne Luo, Ruixiong Zhang, Shuaijiang Zhao, Wubo Li, Cheng Gong, Wei Zou, Kun Han, and 1 others. 2021. Didispeech: A large scale mandarin speech corpus. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6968–6972. IEEE.

Jiarui Hai, Yong Xu, Hao Zhang, Chenxing Li, Helin Wang, Mounya Elhilali, and Dong Yu. 2024. Ezaudio: Enhancing text-to-audio generation with efficient diffusion transformer. *arXiv preprint arXiv:2409.10819*.

Moayed Haji-Ali, Willi Menapace, Aliaksandr Siarohin, Guha Balakrishnan, Sergey Tulyakov, and Vicente Ordonez. 2024. Taming data and transformers for audio generation. *arXiv preprint arXiv:2406.19388*.

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. 2023. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *International Conference on Machine Learning*, pages 13916–13932. PMLR.

Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Amir Ali Bagherzadeh, Chuan Li, Rafael Valle, Bryan Catanzaro, and Soujanya Poria. 2024. Tangoflux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization. *arXiv preprint arXiv:2412.21037*.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. 2018. Fr\'echet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Defossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. 2023. Audiogen: Texturally guided audio generation. In *ICLR*.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rasmus Moritz, Mary Williamson, Vimal Sharma, Yossi Bassett, Yossi Adi, and 1 others. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. In *NeurIPS*.

Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. 2022. Bigvgan: A universal neural vocoder with large-scale training. In *The Eleventh International Conference on Learning Representations*.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. 2023. Audioldm: Text-to-audio generation with latent diffusion models. *ICML*.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. 2024. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883.

Chunyu Qiang, Kang Yin, Xiaopeng Wang, Yuzhe Liang, Jiahui Zhao, Ruibo Fu, Tianrui Wang, Cheng Gong, Chen Zhang, Longbiao Wang, and 1 others. 2025. Instructaudio: Unified speech and music generation with natural language instruction. *arXiv preprint arXiv:2511.18487*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Xu Tan and 1 others. 2024. Naturalspeech 3: Facodec and factorized diffusion for zero-shot tts. *arXiv preprint arXiv:2403.03100*.

Qwen Team and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3).

Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, and 1 others. 2023. Audiobox: Unified audio generation with natural language prompts. *arXiv preprint arXiv:2312.15821*.

Chengyi Wang, Sanyuan Chen, Yu Wu, and 1 others. 2023a. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. 2023b. Cam++: A fast and efficient network for speaker verification using context-aware masking. *arXiv preprint arXiv:2303.00332*.

Jun Wang, Xijuan Zeng, Chunyu Qiang, Ruilong Chen, Shiyao Wang, Le Wang, Wangjing Zhou, Pengfei Cai, Jiahui Zhao, Nan Li, and 1 others. 2025a. Klingfoley: Multimodal diffusion transformer for high-quality video-to-audio generation. *arXiv preprint arXiv:2506.19774*.

Xiaopeng Wang, Chunyu Qiang, Ruibo Fu, Zhengqi Wen, Xuefei Liu, Yukun Liu, Yuzhe Liang, Kang Yin, Yuankun Xie, Heng Xie, and 1 others. 2025b. M3-tts: Multi-modal dit alignment & mel-latent for zero-shot high-fidelity speech synthesis. *arXiv preprint arXiv:2512.04720*.

Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. 2024. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. *arXiv preprint arXiv:2409.00750*.

Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Xixin Wu, and 1 others. 2023. Uniaudio: An audio foundation model toward universal audio generation. *arXiv preprint arXiv:2310.00704*.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*.

Han Zhu, Wei Kang, Zengwei Yao, Liyong Guo, Fangjun Kuang, Zhaoqing Li, Weiji Zhuang, Long Lin, and Daniel Povey. 2025. Zipvoice: Fast and high-quality zero-shot text-to-speech with flow matching. *arXiv preprint arXiv:2506.13053*.

A Related Work

A.1 Text-to-Speech with Voice Cloning

Recent TTS advancements have moved from mel-spectrogram regression to discrete codec modeling and latent diffusion. VALL-E (Wang et al., 2023a) showed that treating TTS as a language modeling task with discrete audio tokens enables zero-shot capabilities. Diffusion-based and flow-matching models like NaturalSpeech (Tan et al., 2024), Voicebox (Le et al., 2023), and CosyVoice (Du et al.,

Data Type	Formatted Instruction Prompt
Type A (Speech)	This audio contains speech. This audio does not contain sound effect. This audio has a speech description: "gender: female, age: elderly, emotion: calm, style: informative, conversational."
Type B (Sound)	This audio does not contain speech. This audio contains sound effect. This audio has a sound effect description: "Before the sound of ceramic plates being placed in a microwave oven, the gentle sound of raindrops can be heard falling softly on an umbrella."
Type C (Mixed)	This audio contains speech. This audio contains sound effect. This audio has a speech description: "gender: female, age: young adult, emotion: happy, style: casual." This audio has a sound effect description: "Before the sound of ceramic plates being placed in a microwave oven, the gentle sound of raindrops can be heard falling softly on an umbrella."

Table 5: Representative samples of structured text instructions from the Multi-Task Compositional Audio Dataset. The prompts are constructed to explicitly define modality presence and provide detailed semantic attributes for both vocals and environmental context.

2024) achieve state-of-the-art fidelity by modeling mel-spectrograms or latent representations. These models typically rely on prompt-based systems with reference audio to dictate timbre, but they excel only in clean speech synthesis and lack control over environmental acoustics, limiting their use in complex auditory scene generation.

A.2 Text-to-Audio Generation

Parallel to TTS, Text-to-Audio (TTA) generation focuses on synthesizing environmental sounds and music from natural language descriptions. The dominance of Latent Diffusion Models (LDMs) is evident in works like AudioLDM (Liu et al., 2023) and Tango (Ghosal et al., 2023), which leverage contrastive language-audio pretraining (CLAP) or large language models (FLAN-T5) to align semantic text embeddings with audio latents. Make-An-Audio (Huang et al., 2023) further explored variable-length generation. Despite their success in generating diverse sound effects, these models struggle significantly with speech generation. Due to the lack of fine-grained phonetic alignment and speaker identity control, the speech produced by TTA models is often unintelligible or hallucinated, creating a distinct capability gap between TTA and TTS systems.

A.3 Unified Audio Generation

Recent research bridges TTS and TTA through unified frameworks. UniAudio (Yang et al., 2023) uses multi-task learning with discrete tokens for speech, music, and sound effects but suffers from autoregressive latency. Audiobox (Vyas et al., 2023) unifies these tasks using flow matching with task-specific guidance. InstructAudio (Qiang et al., 2025), most relevant to our work, employs a multi-modal diffusion transformer for instruction-based audio generation. However, it faces "one-to-many" ambiguity, as textual instructions cannot capture speaker identity, limiting voice cloning fidelity. Additionally, few works tackle compositional speech and sound effects generation within a single model. In contrast, ChameleAudio introduces a hybrid conditioning strategy combining LLM-derived instructions and speaker embeddings, alongside curriculum learning, achieving high-fidelity generation for both separate and mixed tasks, overcoming limitations in previous models.

B Data Samples

Table 5 illustrates specific examples of the structured instructions used in our training pipeline. These samples demonstrate how ChameleAudio differentiates between speech, sound, and mixed modalities using explicit inclusion/exclusion indicators and fine-grained attribute descriptions.